Spring 1-1-2011

Selective Attention as an Example of Building Representations within Reinforcement Learning

Fabian Francisco Canas
University of Colorado at Boulder, canas@colorado.edu

Selective Attention as an Example of Building Representations within Reinforcement Learning

by

Fabián F. Cañas

B.S., Cornell University, 2007

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Arts

Department of Psychology and Neuroscience

2011

This thesis entitled:
Selective Attention as an Example of Building Representations within Reinforcement Learning
written by Fabián F. Cañas
has been approved for the Department of Psychology and Neuroscience

Matt Jones

Prof. Randall C. O'Reilly

Prof. Michael C. Mozer

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB protocol #0208.6

Cañas, Fabián F. (M.A., Cognitive Psychology)

Selective Attention as an Example of Building Representations within Reinforcement Learning

Thesis directed by Prof. Matt Jones

Humans demonstrate an incredible capacity to learn novel tasks in complex dynamic environments. Reinforcement learning (RL) has shown promise as a computational framework for modeling the learning of dynamic tasks in a biologically plausible way. However the learning performance of RL depends critically on the representation of the task. In the machine learning literature, representations are carefully crafted to capture the structure of the task, whereas humans autonomously construct representations during learning. In this work I present a framework integrating RL with psychological mechanisms of representation learning. One model presented here, Q-ALCOVE, explores how RL can adapt selective attention among stimulus dimensions to construct a representations in two different tasks. The model proposes that selective attention can be learned indirectly via internal feedback signals central to RL. I present the results of a behavioral experiment supporting this prediction as well as modeling work suggesting a broad psychological scope for RL.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

The aim of this thesis is to reframe and offer new context to various fields of research such that they may converge to a single unified approach and theoretical framing. In the broadest sense, I claim to solidify a convergence of the machine learning literature with the literature of representation in cognitive psychology. The work so far sets its sights on core principles from each field such such that I hope to present, if not a clear path then at least, an enticing direction forward.

Reinforcement Learning (RL) is a family of machine learning approaches that has become very successful at performing a wide variety of dynamic tasks (Sutton & Barto, 1998). Machine learning approaches using RL methods have accomplished such impressive feats as creating a machine backgammon player better than any human player (Tesauro, 1995), flying a helicopter (Bagnell & Schneider, 2001), and flying a helicopter acrobatically (Ng et al., 2006). But for all their successes, RL learners tend to lack generality of tasks because they have each been specifically engineered to succeed in their target domain. The learners generally begin not with a blank slate, but with a representation of the task provided by human experts familiar with the task and with insights to its solution. These representations describe how to generalize across states in a task, what information is most important, and generally highlight key cues that are critical to success in the task.

In contrast, humans have an incredible capacity for learning a variety of new and complex tasks. Humans may sometimes be presented with a new problem in a similar fashion to these RL learners by being instructed in how to think about a task, how to approach it, and what information to attend to. But interestingly, people can explore and learn new tasks on their own with no

guidance beyond environmental feedback and their autonomously developed representations.

While representation can mean many things, for this work I consider representation to be the structures that underlie similarity. Similarity describes the degree to which two things are or are treated as the same or different. This in turn provides a basis for generalization. Appropriate generalization is the desired end result of good representations as generalization allows a novel situation to be treated by a learner as though it were a familiar situation. Generalization means the learner can use its existing knowledge about similar states to act appropriately in a new and previously unencountered situation.

But representation is more general than the particular mechanisms that facilitates generalization. A representation is the structure inside of which learning takes place. Any learning mechanism requires an environment from which to learn. The general goal is for the learner to anticipate its environment, often in the service of more effectively manipulating its environment. As we try to discover the mechanisms of learning in the brain, it is entirely expected that we will discover mechanisms which are entirely subsumed by the brain–that is, the learner's environment **is** the brain. I propose that Reinforcement Learning is precisely that sort of learning mechanism. The neural correlates of RL do not hold a privileged position to access the outside world exactly as it exists. Rather, the basal ganglia, arguably the root of any instantiations of RL in the brain, and RL as a whole exists as a learning mechanism whose environment is the brain itself. My own line of research intends to elucidate the ways in which RL can act as a general-purpose learning mechanism capable of driving learning in a wide variety of situations, and naturally integrate with other cognitive processes by virtue of being strongly tied to processes that filter and reshape the outside world. By the same expectations that lead us to not expect an RL learner to receive raw sensory information, we should expect that well-established cognitive mechanisms for reshaping representation to have the capacity to act **prior to, and in tandem with** learning, and not unaccountably separately from it.

My interest in the convergence of RL and cognitive representation mechanisms is two-fold: it has a clear capacity to make machine learners more flexible by incorporating elements of human

representation learning into RL and describing mechanisms by which an RL model can change its own representation. This, in turn, will inform the development of psychological theories of learning, and help us to closely examine the parallels between biological analogs of RL and their mathematical counterparts.

The psychological impact of this line of research stems out of further exploration of the scope of different forms of learning. Mechanisms for psychological representation learning have thus far tended to remain in the domain of categorization or stimulus discrimination, where the principles were discovered and investigated. A natural extension of the present work is to investigate the properties of such representation mechanisms when learned in, and used in the service of, dynamic tasks. Additionally, work with human subjects performing dynamic tasks will help define the scope of RL in psychological contexts. An open question in psychological RL is whether the simple but powerful feedback signals of RL actually act in the service of learning complex tasks, or are relegated to simpler learning along the lines of operant conditioning (Fu & Anderson, 2008). This thesis explores the possibility that RL is tightly integrated with psychological mechanisms of representation.

In this thesis, I present modeling evidence contrasted with existing experimental work to argue that Reinforcement Learning is appropriately suited to interact with other cognitive processes as well as account for more complex behavior than might be initially supposed (Chapter 4). I present a formal model incorporating principles of human selective attention learning with the powerful learning mechanisms of RL (Chapter 5). The model is intended to demonstrate that useful representations can be constructed through the updating signals at the heart of RL, and that the learner in turn benefits from a flexible representation in the way of faster learning. Though the proposed model integrates two psychologically well-supported mechanisms, their combination is novel, and so I will also present the results of a human experiment (Chapter 6) and additional modeling work to validate the combination (Chapter 7). The experiment aims to validate the novel psychological assumption that arises from the combination of the two existing models, that RL can drive representation learning.

Chapter 2

Reinforcement Learning

RL is a computational framework for learning optimal actions in dynamic tasks. A dynamic task is one where actions taken in an environment can have an impact on future states encountered in that environment. RL represents dynamic tasks as a set of states, together with a set of actions available at each state. The selection of any action determines the immediate reward as well as the ensuing state. This general framework of describing a task encompasses most interesting real-world tasks, which are usually temporally extended and involve interaction with an environment. This can be contrasted with classical learning paradigms, such as those used when studying category learning, where interactions take the form of stimulus-response, but there is no interaction with the processes responsible for generating the stimuli.

RL's approach to learning dynamic tasks is to estimate **values** of states and actions, which reflect predictions of the reward the learner can expect from all future states, beginning with the state and action whose value is in question. From any given state, the action with the highest estimated value represents a best guess of the choice that will lead to the highest long-term reward. The key to learning value estimates is an internally generated feedback signal known as Temporal Difference (TD) error. TD error represents the discrepancy between the estimated value of an action prior to its execution and a new estimate based on the immediate reward and the value of the ensuing state. Within the machine learning literature, TD is used to refer to a specific RL algorithm that associates values to states. The broader sense of TD as it is used here is consistent with its use within the psychological literature.

Formalization starts by defining a value, V, as

$$V(s) = E\left[\sum_{k\geq 0} \gamma^k r_{t+k} | s = s_t\right], \tag{2.1}$$

where V represents the expected total future discounted reward for state s at time t. The reward from the environment at time t is r_t, and $\gamma \in [0, 1)$ is a temporal discount factor which represents the relative value placed on temporally proximal versus distant reward. We next derive a recursive definition of V_t (Bellman, 1957):

$$V(s) = E\left[\sum_{k\geq 0} \gamma^k r_{t+k} | s = s_t\right] \tag{2.2}$$

$$= E\left[r_t + \gamma \sum_{k\geq 1} \gamma^{k-1} r_{t+k} | s = s_t\right] \tag{2.3}$$

$$= E\big[r_t | s = s_t\big] + \gamma E\left[\sum_{k\geq 0} \gamma^k r_{(t+1)+k} | s = s_t\right] \tag{2.4}$$

$$= E\big[r_t | s = s_t\big] + \gamma E\big[V(s_{t+1}) | s = s_t\big] \tag{2.5}$$

$$V(s) = E\big[r_t + \gamma V(s_{t+1}) | s = s_t\big] \tag{2.6}$$

$V(s)$ is is now re-expressed in terms of the value of the ensuing state $V(s_{t+1})$ and the immediate reward, r_t. This leads to a class of algorithms for learning these values. At every transition between states, the accuracy of the prediction can be evaluated with information that is immediately available. The discrepancy between the predicted value and the observed instance of what the value is predicting is defined as the TD error:

$$\delta = \big(r_t + \gamma \cdot V(s_{t+1})\big) - V(s_t). \tag{2.7}$$

If we consider Equation 2.7 from a learning perspective, once all learning has taken place, the rightmost part of the equation, $V(s_t)$, should be equal to the leftmost part of the right-hand side, $r_t + \gamma \cdot V(s_{t+1})$ and δ becomes zero. Thus TD error can be used as a driving force for learning by adjusting V to reduce error after every action, scaled by a learning rate $\epsilon \in [0, 1]$:

$$\Delta V(s_t) = \epsilon\delta. \tag{2.8}$$

TD error is centrally important to RL, and its general principles and structure are referenced throughout this work.

2.1 Neural Correlates

The importance of RL extends beyond its normative basis and proven power in the realm of machine learning. Psychologically, RL represents a generalization of the Rescorla-Wagner model of error-driven learning (Rescorla & Wagner, 1972) to accommodate error propagating through time (Sutton & Barto, 1990). In the last twenty years there has been mounting evidence indicating that portions of the basal ganglia are producing the most important computational characteristics of TD error. Dopaminergic neurons in the basal ganglia respond to reward (an unconditioned stimulus) during learning, but not after an associated learned behavior has been established (Schultz, Apicella, & Ljungberg, 1993; Schultz, Dayan, & Montague, 1997), in correspondence to TD error in response to an unexpected reward. DA neurons' firing then temporally shifts to associate with a conditioned stimulus that occurs before the reward, corresponding with the conditioned stimulus taking on an internal value in anticipation of future reward (Schultz et al., 1997; Schultz, 1998). And finally, when a reward anticipated by a conditioned stimulus is not present DA firing drops below baseline at the time of expected reward, corresponding with a negative temporal difference error (Schultz et al., 1997; Schultz, 1998).

Though it has been suggested that the TD signals computed in the basal ganglia act primarily to drive motor (Holroyd & Coles, 2002) and procedural learning (Fu & Anderson, 2008), human and animal evidence shows that cognitive functions such as rule-based learning suffer as a result of BG damage (Doya, 2000). Extensive connections between the basal ganglia and cortical areas (Alexander, DeLong, & Strick, 1986) suggests that BG activity may be acting on more elaborate cognitive processes (Frank, Loughry, & O'REILLY, 2001).

2.2 Control, Policy and Q-Learning

The methods in the opening of this chapter outline how values for states can be estimated from experience. What they do not do is provide any method for determining what action to take. $V(s_t)$ provides no information about determining an action selection policy. What it does provide is an estimate for the value of being in a given state assuming the policy in effect during learning continues to be in effect. To effectively develop methods of control within the RL framework, we must extend the concept of $V(s_t)$ as a value for a given state, by also accounting for actions that can be taken from that state. The **quality** of taking an action a_t from a state s_t is denoted $Q(s_t, a)$ and is termed a Q-value. Where each state has only one state value estimate $V(s)$, each state has as many Q-values as there are actions available to an agent in that state.

A single Q-value, $Q(s_t, a_t)$, represents the expected total future reward for choosing action a_t from state s_t. One algorithm that uses Q-values, SARSA (Rummery & Niranjan, 1994), updates Q-values based not only on the ensuing state, but the action taken from there as well, leading to a slightly modified TD error in Equation 2.9,

$$\delta = r_t + \gamma \cdot Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t), \tag{2.9}$$

and essentially the same learning rule as in Equation 2.8, but updating Q and using the new δ.

Since SARSA uses the actual results of the actions it takes to adjust the Q-values, it is not guaranteed to converge on an optimal set of Q-values without a proper exploration, primarily requiring that the policy becomes greedy in the limit and also explores infinitely (GLIE) (Singh, Jaakkola, Littman, & Szepesvári, 2000). As an example of an SARSA learner whose Q-values will not converge, consider an agent that always takes consistently depending on the state, but without considering the actions' predicted value. Eventually, we expect the SARSA learning rule to produce Qs that very accurately predict the value of action the learner actually takes. But the collection of all Qs will not accurately reflect the best actions to take from each state. A different agent with a different policy would not benefit from those values. This property makes SARSA an on-policy learner, in that the Q-values it converges to reflect the expected rewards following a given policy,

much in the same way as state value estimates $V(s)$ outlined in the previous section.

On-policy algorithms like SARSA can be contrasted with off-policy techniques where Q-values are updated without regard to the choices the current policy is making. An off policy algorithm which will be referenced and extended throughout the remainder of this thesis is Q-learning (Watkins & Dayan, 1992). Q-learning updates its Q-values so that they reflect the best known possible action for the task, regardless of the action-selection policy the learner is using. This is done by reformulating the TD error, replacing the reference to the action actually taken at $t+1$ with the best known action at s_{t+1}, as in Equation 2.10.

$$\delta = r_t + \gamma \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \tag{2.10}$$

This approach affords the learner the flexibility to not choose the optimal action at for s_{t+1}, that is to **explore**, and not lose accuracy in its estimated maximum value, $Q(s_t, a_t)$. If the learner always selects the action with the maximum Q before the Q-values have settled from early on in learning where Q values don't accurately reflect the structure of the task, it risks settling on non-optimal state-action value estimates and not learning anymore while at the same time performing poorly. This is a problem of exploration versus exploitation. Where a policy of exploitation will take the action with the highest predicted value every time, a policy of exploration will sometimes not. Exploration is helpful, especially early in the learning process, because the set of Q-values may not reflect what they should optimally be. It is important for the learner to try all possible actions from within a state to accurately estimate their values.

A simple selection policy addressing the issue of exploration, ϵ-greedy, chooses the best action for a proportion of $1 - \epsilon$ of trials. On the remaining trials, the other actions are chosen at random (Sutton & Barto, 1998). This allows for varying levels of exploration of the state space, depending on the value of ϵ. Another strategy allowing for exploration, and the strategy I use later in this proposal, is the softmax function:

$$p_i = \frac{\exp(Q(a_i)/\tau)}{\sum_j \exp(Q(a_j)/\tau)}, \tag{2.11}$$

where p_i is the probability of selecting the ith action, and τ is a temperature parameter. Softmax allows for exploration while favoring actions with high Q-values. Actions with larger Q-values have a higher probability of being selected. Lower temperatures lead to the highest Q-value being selected more often. Higher temperatures lead to a more exploratory policy.

2.3 RL and Generalization

A critical question for all RL models concerns the relationship between value estimates (Q or V) for different states. The various learning rules presented above suggest that separate values can be kept for every state. But for most realistic tasks with large state spaces and high dimensionality this is unfeasible. Effective learning requires **generalization**, or the use of knowledge about one state to make inferences or choose actions for other, similar states. A number of methods have been proposed for implementing generalization in RL, such as radial basis functions (Poggio & Girosi, 1998), and tile coding (Albus, 1981). However, the pattern of generalization depends strongly on the way in which states are represented.

An illustrative example in the importance of representation to RL can be seen in an advanced backgammon player, TD-gammon (Tesauro, 1995). TD-gammon works by encoding each board configuration as a set of values along a set of feature dimensions. These features act as the input to a neural network whose job it is to learn action values to associate with its input. But the features are not a simple enumeration of piece position. Instead they were hand-coded by experts so that board representation reflected important elements in the game. Generalization then becomes easy for TD-gammon because two board configurations could be similar in representation with respect to the features they are encoded by, while being superficially different to an untrained player. This choice of features positively affected the pattern of generalization among states. Representations relying on different features produce different patterns of similarity and hence different generalization. Learning will be most effective if generalization respects the structure of the task, such that the learner pools knowledge about states with similar consequences but discriminates between states that are meaningfully different.

Chapter 3

Representation Learning

The most flexible learner will be one who can shift representations to match what is known about the task at hand as new information becomes available. Humans learning a new task may be instructed in how to think about and approach a new problem; this is analogous to an RL learner being given a pre-defined representation. However humans have the additional ability to alter these representations, or to construct entirely new ones through mechanisms which include selective attention (Kruschke, 1992; Nosofsky, 1986); feature discovery (Schyns, Goldstone, & Thibaut, 1998), prototype formation (Smith & Minda, 1998); hybrid rule-exemplar representations (Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994); clustering representations that grow with task complexity (Anderson, 1991); mutable representations that evolve among exemplars, prototypes, and rules (Love & Jones, 2006; Love, Medin, & Gureckis, 2004).

In this work, I focus on the interaction of RL and selective attention among stimulus dimensions. In particular, I will be demonstrating a method for, and the implications of learning selective attention by way of Reinforcement Learning's Temporal Difference error. But it should be clear that the same approach taken in integrating RL with selective attention might be taken with any of the systems of representation mentioned so far, as well as with any mechanisms yet to be described.

3.1 Selective Attention Among Stimulus Dimensions

Though attention has been studied under many guises in psychology, its implications for learning and generalization have been primarily explored in animal conditioning and categorization. In these literatures, attention has been proposed to act by reshaping generalization gradients (Sutherland & Mackintosh, 1971; Nosofsky, 1986). The generalization gradient is an empirical function that describes how strongly subjects generalize between stimuli as a function of how much those stimuli differ, where a greater difference always leads to less generalization. This function decreases more rapidly along attended dimensions than unattended dimensions (Jones, Maddox, & Love, 2005). Thus subjects generalize less between stimuli when they are attending to the dimensions those stimuli differ on. An alternative and equivalent view is that the generalization gradient is fixed and isotropic, but the perceptual scaling of individual stimulus dimensions is adjustable. Attention to a dimension serves to stretch the perceptual space so that stimuli differing on that dimension become less similar and thus produce less generalization (Nosofsky, 1986).

The mechanism of generalization from categorization that I will be adapting is best approached by examining exemplar models (Medin & Schaffer, 1978), where a measure of similarity is made explicit and easy to manipulate. In these models, the psychological evidence (E) in favor of classifying a stimulus (s) into a given category (c) is given by summing its similarity to all previously encountered exemplars (S), weighting each exemplar by its association to c.

$$E(s,c) = \sum_{S} sim(s,S) \cdot w(S,c) \tag{3.1}$$

Additionally, this similarity function has been shown to change during the course of learning as a consequence of shifts of attention among the stimulus dimensions (Nosofsky, 1986). This flexibility is modeled by expressing similarity as a decreasing function of distance in psychological space, with each stimulus dimension, i, scaled by an attention weight, α (Nosofsky, 1986). Here we assume an exponential similarity-distance function, in accord with empirical evidence (Shepard, 1987).

$$sim(s, S) = exp(-\sum_i \alpha_i |s_i - S_i|) \tag{3.2}$$

The influence of attention on generalization has extensive support, both theoretically (Medin, Goldstone, & Gentner, 1993) and empirically (Jones et al., 2005; Nosofsky, 1986). An important question suggested by this research is how attention can be learned. One proposal is that attention weights are updated in response to prediction error (Kruschke, 1992). In a classification task, prediction error (δ) is simply the difference between the category evidence, $E(s, c)$, and the actual category membership given as feedback to the learner (e.g., $+1$ if $s \in c$ and -1 otherwise). The updating rule for attention is then based on gradient descent on this error, squared and summed over categories.

$$\Delta\alpha = -\epsilon_{att} \cdot \frac{\partial}{\partial \alpha_i} \left(\frac{1}{2} \sum_c \delta_c^2 \right) \tag{3.3}$$

This mechanism for attention learning has been implemented in ALCOVE, a highly successful model of human category learning (Kruschke, 1992). ALCOVE learns to shift attention to stimulus dimensions that are most relevant to predicting category membership and away from dimensions that are non-diagnostic. This leads to adaptation of generalization, which in turn speeds learning.

Chapter 4

Reinforcement Learning and Memory

While the neural correlates of RL have a strong foundation, as described in Section 2.1, the extent of RL's influence on higher-level cognitive function remains unclear. Fu and Anderson (2008) investigated the possibility of two separate learning systems in a dynamic task; one learning system was argued to be RL-like, while the other distinctly different and dependent not on temporal difference errors but on semantic memory.

One basic feature of many simple RL models is that many passes through a task are often required to fully learn it. This is because when a reward is encountered, only the value of the state immediately preceding the reward is updated to reflect that reward. For the information about that reward to propagate to earlier in the task, those same states and actions must be experienced again. So each pass though a task allows information to propagate one step closer to the beginning of the task. While this basic pattern is appropriate for modeling procedural learning, it does not always reflect how humans learn other tasks. Under fairly simple circumstances humans may learn to perform the first step of a task before later step (Fu & Anderson, 2008).

Fu and Anderson (2008) suggest that RL only operates in circumstances where learning is implicit, and a separate explicit learning system predominates otherwise. They cite the dissociation in performance between amnesic patients and patients with disorders of basal ganglia function as their motivating evidence. Where amnesic patients can learn some tasks well-suited to RL while having no declarative access to their strategy or training, patients with basal ganglia disorders cannot perform those tasks. However the dissociation between learning in amnesiac patients versus

those with basal ganglia disorders is not complete. It is not clear whether disorders of the basal ganglia spare declarative learning mechanisms. This leaves open the possibility that the basal ganglia are important for both procedural as well as semantic learning. Fu and Anderson intend to relegate biological RL to low-level procedural learning, and suggest that any learning involving more complex representations relies on a different learning mechanism. I argue that behavioral dissociations and the further evidence from a behavioral experiment (Fu & Anderson, 2008) can be explained by a single **learning** mechanism based on RL, and that the introduction of a memory system can account for behavioral differences without the need for parallel learning systems.

To investigate the possibility of two parallel learning systems, Fu and Anderson (2008) trained and tested subjects on a dynamic task, and across two groups manipulated the demand on subjects' working memory. The experimental group learned and performed a probabilistic navigation task, described below, while performing a secondary task designed to impair their working memory. The control group performed the navigation task with no additional demands. Fu and Anderson suggest that the additional task demands on the experimental group act as a switch between the learning systems. Here, I describe a learning system based on Reinforcement Learning that can account for the differences in behavior between the groups by manipulating a 'memory' parameter instead of relying on two independent learning systems. This suggests that the demands on subjects' working memory does not function as a switch between learning systems, but impairs a memory system that supplements a single common learning mechanism.

Fu and Anderson's (2008) probabilistic rooms task, shown in Figure 4.1, consisted of three states, each with two actions available. Actions are coded by color, and the states, though referred to by number are only encoded by the names of arbitrary objects to the subjects (**e.g** "Books", or "Computers"). Shown in the figure are transition probabilities associated with each action. Reaching the exit is the desired outcome, where not reaching it is undesirable.

A key feature of the experiment is that the probability of reaching the exit is much higher from from Room 2 than from Room 3. In particular, subjects acting randomly on Step 2 still have performance in the task depend on the actions at Step 1. Notice too that the certainty of the

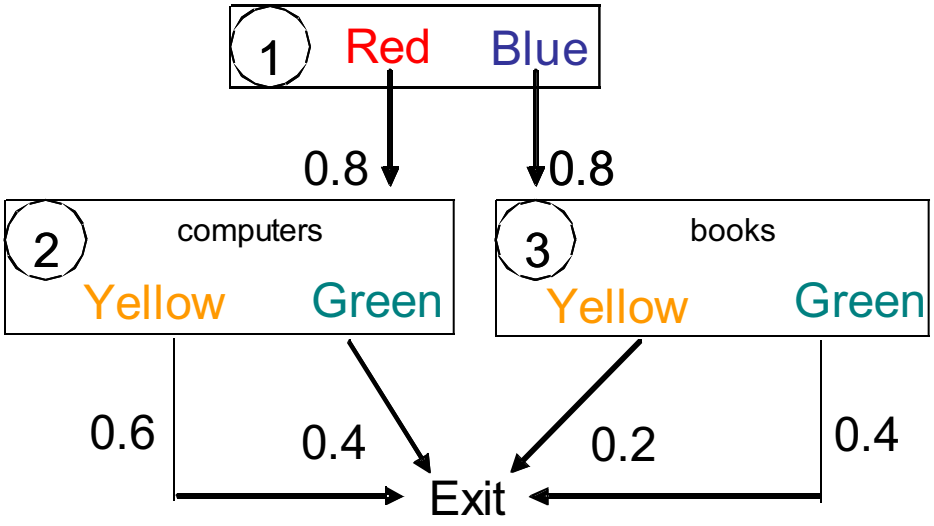

Figure 4.1: Outline of probabilistic rooms task reproduced from Fu and Anderson (2008). Actions are coded by color, and the states, though referred to by number are only encoded by the names of arbitrary objects to the subjects (**e.g** "Books", or "Computers"). Numbers on the arrows between states are transition probabilities associated with each action. The inverse of these numbers is the probability of arriving at the **other** state for that step of the task. Reaching the exit is the desired outcome, where not reaching it is undesirable and an implicit additional state at the bottom of the diagram..

outcome on Step 1 is much higher than in Step 2. This set of conditions is responsible for a key finding that subjects subjects not burdened by an additional demand on memory learn the first step of the task before they learn the second. This appears to contradict a general prediction of Reinforcement Learning: that later stages of a task are learned first, and information propagates towards the start of a task through repeated experience.

However, a simple RL model augmented with a simple working memory system can account for the switching of learning order in the task. A model with an intact memory can learn the first step of the task first, and the same model with its memory impaired will learn the second step first.

4.1 Eligibility Traces

An eligibility trace is a vector of weights for all states representing their current activation level. The weight for each state represents the degree to which that state is "eligible" for learning. States that are eligible receive some amount of credit for actions taken **after** those states are encountered, and their activation, or eligibility, may take many time steps to decay. This is in contrast to simpler RL models which only update the last encountered state. Updating earlier states provides advantages in the form of potentially faster learning, and an improved ability to learn tasks where the information available to the learner does not fully specify the state of the environment (Jaakkola, Singh, & Jordan, 1995; Singh, Jaakkola, & Jordan, 1994). It may also impair learning by assigning undue credit to states distant and unrelated to a reward. Each encountered state becomes activated within the eligibility trace, and that activity decays as the task proceeds. The eligibility trace therefore acts as a memory of recent states, and values for states continue to be updated for as long as they have corresponding activity in the eligibility trace. Since the contents of an eligibility trace are dictated by the recency of encountered states, it has been proposed to act in a manner roughly analogous to working memory in RL models (Holroyd & Coles, 2002). Eligibility traces, as presented here, are intended to serve as the simplest model of working memory that gives rise to the pattern of results observed by Fu and Anderson (2008), and not as a complete model of the relationship between RL and working memory.

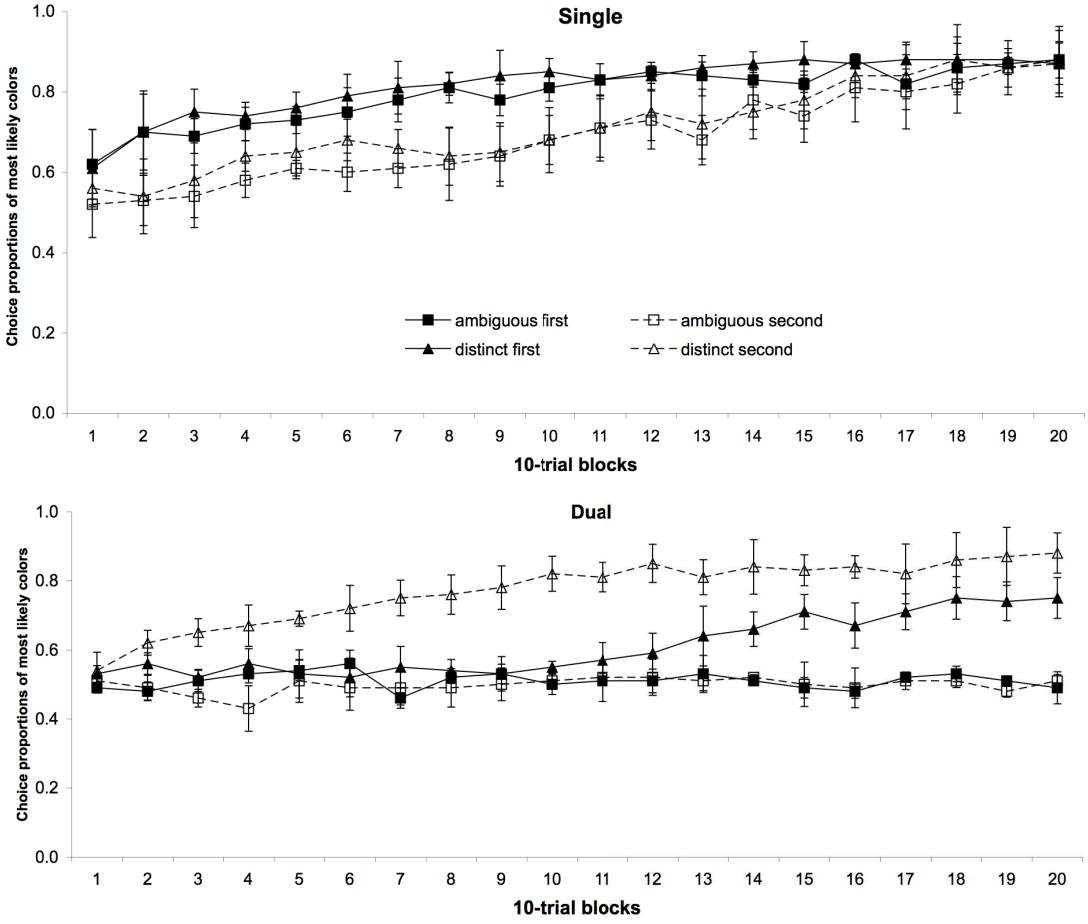

Figure 4.2: Performance curves from Fu and Anderson (2008), to be compared with computational modeling results in Figure 4.3. The first subfigure shows the single-task condition, which should be compared to $Q(\lambda)$. The second subfigure shows the dual-task condition, which should be compared to tabular Q-learning. Only the distinct conditions (triangles) were modeled. Notice that in the dual-task condition, the second step is learned before the first, whereas in the control condition the first is learned before the second. A discussion of modeling the ambiguous conditions (squares) can be found in Section 4.3.

An eligibility trace defines a weighting by which an internal value estimate, **e.g.** $V(s)$ or $Q(s,a)$, is updated during learning. Each value estimate has a corresponding weight in an eligibility trace, η, and so η is indexed as the value is indexed. For an eligibility trace on Q-values, the trace is indexed $\eta(s,a)$ and has the an update rule on each time step shown in Equation 4.1.

$$\eta_t(s,a) = \begin{cases} \gamma\lambda\eta_{t-1}(s,a) + 1 & \text{if } s = s_t \text{ and } a = a_t; \\ \gamma\lambda\eta_{t-1}(s,a) & \text{otherwise} \end{cases} \tag{4.1}$$

The contents of the eligibility trace are controlled by the exponential decay parameter, $\lambda \in [0,1]$. A model with $\lambda = 0$ has no memory and is equivalent to whichever RL model it is based on, with no eligibility trace. Only the previously encountered state is updated. A model with $\lambda = 1$ has a perfect memory, and all states that are encountered are available to receive full and equal credit for all actions that are taken. Intermediate values of λ provide an exponentially decaying trace, where recently encountered states are updated more than states encountered further in the past. The same temporal discount parameter, λ, used in standard TD equations (Equation2.7).

Though eligibility traces are compatible with any RL method, for simplicity I consider only variants of Q-learning with eligibility traces; there have been at least two such combinations. The first, Watkins' (1989) $Q(\lambda)$, is a more formal combination of Monte Carlo methods with TD. It takes a series of greedy actions, building an eligibility trace as it goes, not updating according to that trace but instead updating according to standard Q-learning, which updates only the previous state. On the first non-greedy action that is taken, the state-action pairs in the eligibility trace are updated. Before the following step, the eligibility trace is cleared. If an eligibility trace is to serve as an analog to working memory, one would not expect for a subjects' memory to clear based on whether an action that was taken is exploratory or not. Other common formulations of eligibility traces do not frequently clear those traces, and so are more appropriate here.

Another combination of Q-learning with eligibility traces is Peng's $Q(\lambda)$ (Peng & Williams, 1996). The approach's key feature is that it calculates two separate TD-errors: one

$$\delta = r + \gamma \max_{a'} Q(s',a') - Q(s,a) \tag{4.2}$$

is used for updating $Q(s_t, a_t)$, just as in Q-learning, and another

$$\delta' = r + \gamma \max_{a'} Q(s', a') - \max_{a'} Q(s, a) \tag{4.3}$$

for updating all Qs via their presence in the eligibility trace. Peng's $Q(\lambda)$ maintains an eligibility trace throughout learning, and also makes efforts towards keeping $Q(\lambda)$ an off-policy learner by calculating a special TD error for the eligibility trace comparing the maximum value of the next state (consistent with Q-learning) against the maximum value of the current state instead of the expected value (Eq 4.3). It is important to consider, though, that the off-policy nature of Q-learning must almost necessarily be violated with the incorporation of eligibility traces, as traces are only constructed by past actions, and are therefore entirely dependent on the current policy.

To maintain simplicity in the following demonstration, a naive version of $Q(\lambda)$ was implemented. It neither clears its eligibility trace after non-greedy actions, nor does it calculate two different TD-errors at every step. For any given state, s, the learner chooses an action according to a softmax selection rule on the Q-values of the available actions (Equation 2.11). A temporal difference error is calculated just as in simple Q-learning in Equation 4.2. The eligibility trace $\eta(s, a)$ is is updated on each step t according to Equation 4.1. All Q-values are updated at each step in proportion to their corresponding eligibility value in η as defined in Equation 4.4. The result is that the most recently visited state is guaranteed to be updated, and states that have not been visited are not updated at all. States that have been recently visited are updated in accordance to how recently they were visited as dictated by the trace decay parameter $\lambda \in [0, 1]$. When $\lambda = 0$, we have a special case where the algorithm is tabular Q-learning because the previous step has a eligibility of one and therefore is updated normally, while all prior steps have had their eligibility set to zero. When $\lambda = 1$, we have a learner with a perfect memory because the eligibility of all previously encountered states never decays, and all previously encountered states are updated (see Equations 4.1 and 4.4).

$$Q_{t+1}(s, a) = Q_t(s, a) + \epsilon \cdot \delta \cdot \eta_t(s, a) \tag{4.4}$$

Algorithm 4.1 Naive $Q(\lambda)$

Initialize s, $Q(S, A) \leftarrow 0$, $\eta(S, A) \leftarrow 0$
Repeat for each step in an episode:
  Select action $a \in A_s$ according to softmax and $Q(s, A_s)$ (Eq. 2.11)
  Take action a; observe reward r and next state s'
  $\delta = r + \gamma \max_{a'} Q(s', a') - Q(s, a)$
  $\eta(s, a) \leftarrow \eta(s, a) + 1$
  $Q(S, A) \leftarrow Q(S, A) + \epsilon \cdot \delta \cdot \eta(S, A)$
  $\eta(S, A) \leftarrow \gamma\lambda\eta(S, A)$
  $s \leftarrow s'$
$\eta(S, A) \leftarrow 0$

The full description of Naive $Q(\lambda)$ is given in Algorithm 4.1. It is important to note that for the task examined in Section 4.2, the eligibility trace is cleared between episodes as shown in the last line in Algorithm 4.1. This was done because it was made clear to subjects in the behavioral task that the episodes were independent. The model presumes that the strategy of the subjects was to utilize their working memory, modeled by η, **within** a trial to create knowledge about the structure of the task, modeled by Q, that accumulates over episodes.

4.2 Probabilistic Rooms Task

Naive $Q(\lambda)$ was run on a probabilistic room navigation task under two conditions, $\lambda = 1$ models the control condition for the subjects, where the model retains a perfect memory of the first action taken upon receiving reward after the second step. Letting $\lambda = 0$ reduces the model to a tabular Q-learning with no memory and is intended to model subjects with an occupied working memory.

Figure 4.3 shows a comparison of Q-learning and $Q(\lambda)$ in correct action on the first and second steps. $Q(\lambda)$ exhibits the same pattern of results as do subjects not performing an additional memory task, and Q-learning shows the same patten of results as subjects with the additional memory task. $Q(\lambda)$ learns the first step of the task before it learns the second step of the task, while Q-learning learns the second step before the first. Both models are using fundamentally the same learning mechanism based on a temporal difference error. They differ in a single memory

parameter, thus demonstrating that such a pattern alone does not necessitate separate implicit and explicit learning systems.

Figure 4.3 can be contrasted with the actual learning curves from the corresponding empirical study (Figure 4). Subjects' learning curves are flatter, while both $Q(\lambda)$ and Q-learning's curves are more strongly concave and the models learn the task significantly faster. For the purposes of this demonstration, I do not intend $Q(\lambda)$ to serve as a full psychological model, and so I only found parameters that exhibited the same qualitative pattern of learning as subjects. The key parallel is the interaction of which step is learned first with memory demand. It is entirely possible that one of the other two variants of $Q(\lambda)$, or that $SARSA(\lambda)$ (Rummery & Niranjan, 1994), another RL model with eligibility traces would serve as a more complete psychological model. A simple $Q(\lambda)$ was used here only to show that a learning approach centrally dependent on temporal difference error is capable of exhibiting behavior not typically expected of RL models.

Fu and Anderson (2008) ascribe the difference in learning order to the design of their task, as I also do. But they conclude that two learning systems are needed to account for the data. In particular, they aim to eliminate the role TD-error plays in learning when a memory system is fully functional. $Q(\lambda)$ is still a model with two components, an RL component as well as a memory component. I show that it is not necessary to have an entirely different **learning** system in parallel to RL; TD-error, and by association the basal ganglia, may still be playing a critical role in learning tasks that are not procedural but that require other cognitive facilities. As I initially outlined in Chapter 1, if there is to be some form of reinforcement learning in the brain, we should expect that its environment is the brain itself. It would not have privileged access to the raw state of the environment external to the organism, but could use the contents of memory or rich representational structures provided by other brain areas. A rich interaction between RL mechanisms and other cognitive facilities can give rise to complex learning behavior not exhibited by those systems independently.

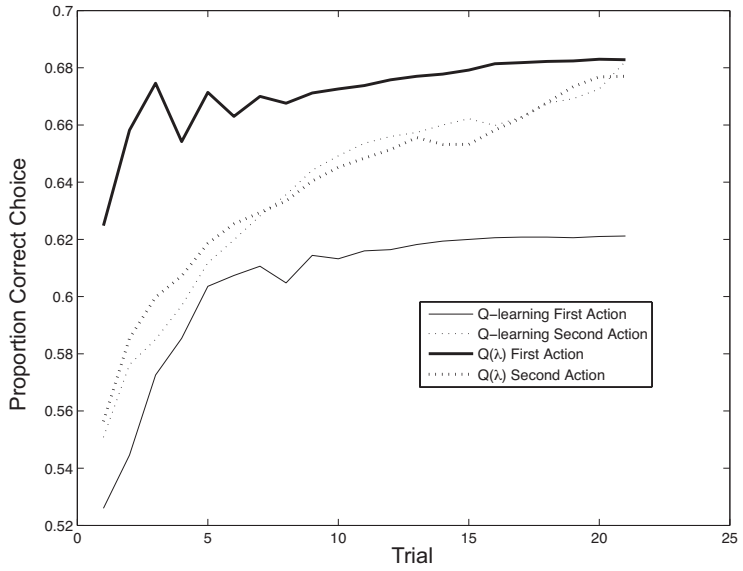

Figure 4.3: Learning curves for the first and second step in a probabilistic room navigation task for two models: $Q(\lambda)$, an RL model where an eligibility trace functions as a working memory; Standard Q-learning. Learning curves are averaged across 1000 independent runs of each model and averaged over a rectangular window of five trials. This averaging window accounts for performance not starting at chance (0.5), as the first point in the plot is the average proportion of correct actions in the first five trials.

4.3 Reinforcement Learning and Partially Observable Markov Decision Processes

Fu and Anderson's experiment had an additional set of conditions where the first or second steps contained no information about the state, labels for the rooms were omitted. For example, after taking an action from the first step, subjects entered one of two different rooms in the second step where different actions are optimal, but had no visible indication as to which room they were in. This is a case where the task is a partially observable Markov decision process (POMDP), as information from prior states or actions is necessary for determining the current state. Just as with the reversed order of learning in the previous section, the findings for subjects performing a task impairing their declarative memory matched what would be expected from simple RL algorithms described above, that steps later in the task are learned first. The pattern of findings for the unimpaired subjects also remained unchanged. Since there is no state information about the second step, most standard RL models cannot distinguish the two possible states and will therefore be unable to learn an optimal action for that step of the task (Jaakkola et al., 1995; Singh et al., 1994).

For one possible method of RL overcoming lack of state information in a dynamic task, reconsider the proposal that if RL were to be implemented in the brain, its environment would be the rest of the brain. It then follows that the contents of the eligibility trace can be treated as a part the current state. The trace would contain states that were previously encountered as well as information about the actions taken from those states, just as in naive $Q(\lambda)$. Psychologically, this is equivalent to the contents of working memory being just as accessible and just as much a part of the present environment as is sensory input. This differs from naive $Q(\lambda)$ in that states are no longer solely explicitly encoded, but are at least partially a function of the eligibility trace. When reliable information is present in the environment, the current state is dominated by sensory input. But when information is unreliable, ambiguous, or a piece of information from memory is deemed important, the identity of the present state is instead dominated by the contents of working

memory. Approaches using $SARSA(\lambda)$ with similar state representations have proven to perform as well on POMDP tasks as models with elaborate methods of state estimation (Loch & Singh, 1998). A dopamine-controlled gating architecture (O'Reilly & Frank, 2006) has recently been used to learn POMDP tasks with excellent success and promising psychological validity (Todd, Niv, & Cohen, 2009).

Chapter 5

Q-ALCOVE

Next, I describe a model that results from a natural synthesis of the RL mechanisms from Q-learning (Chapter 2; Watkins & Dayan, 1992) with the exemplar model of categorization ALCOVE (Chapter 3; Kruschke, 1992). Q-ALCOVE is an RL model that uses selective attention to alter similarity and generalization among states. By virtue of having a mutable measure of similarity, the model can discover optimal patterns of generalization and successfully learn tasks without relying on tailored representations. RL and ALCOVE have been previously combined to make ALCOVE more physiologically plausible (Phillips & Noelle, 2004). Q-ALCOVE differs significantly in that it uses selective attention in a temporally extended task, and not just in stimulus categorization. Q-ALCOVE is shown to learn attention without feedback after every decision.

5.1 Model Specification

Since states are the object of learning for RL, ALCOVE's exemplars represent previously encountered states in Q-ALCOVE. Q-ALCOVE estimates action values by generalizing among neighboring states (Equation 3.1) using ALCOVE's similarity function (Equation 3.2) with the addition of normalization.

$$Q(s,a) = \frac{\sum_{s'}(sim(s,s') \cdot w(s',a))}{\sum_{s'} sim(s,s')} \tag{5.1}$$

The Q-values are used to determine the action selected by the model according to the softmax function (Equation 2.11). Once an action is taken, the learner calculates the TD error, δ, according

to Q-learning (Equation 2.10) and updates the pre-generalization action values, w in ALCOVE where they represent pre-generalization category membership, according to Equation 5.2.

$$\Delta w(s, a_t) = \epsilon_w \cdot \delta \cdot sim(s, s_t) \tag{5.2}$$

The attention learning mechanism from ALCOVE (Equation 3.3) is also modified to use the TD error, δ, and works by gradient descent with respect to α thus yielding Equation 5.3.

$$\Delta \alpha_i = \epsilon_\alpha \cdot \delta \cdot \frac{\partial}{\partial \alpha_i} Q(s_t, a_t) \tag{5.3}$$

ALCOVE itself does not specify how its exemplars are initialized. Q-learning on the other hand, specifies that the full set of Q-values are initialized arbitrarily before any learning takes place. For a tabular learner such as standard Q-learning, arbitrarily initialization of Q-values at the start of learning is no different than initializing a Q on first encountering a state. When we introduce the ability to draw information from other states, however, preallocating all possible exemplars in a task and initializing their values for w means that early on in learning, the learner may be drawing on meaningless information from similar states that have never been visited before. Instead, Q-ALCOVE begins with no exemplars and recruits an exemplar for each novel state it encounters. In practice, both the pre-allocation and the recruitment strategy lead to appropriate behavior; the exemplar-recruitment strategy seems to have better behavior, at least early in learning. The remainder of this work uses a recruitment strategy as shown in Algorithm 5.1.

Q-ALCOVE's learning rules for w and α both critically depend on the TD error, δ. As the model adjusts its pre-generalization action values in Equation 5.2, it is accumulating further knowledge about its current state, as well as all other states in proportion to how much they contributed to the last action, so that the model might more accurately predict the value of the encountered state in the future. α is updated in Equation 5.3 to alter similarity (Equation 3.1) so that exemplar states that contributed more to error become less similar to the current state, and states that contributed to a more correct answer become more similar. Since these similarity

Algorithm 5.1 Q-ALCOVE

Initialize s
$a \leftarrow$ random action from A_s
Take action a; observe reward r and next state s'
$w(s, A) \leftarrow r$ Initialize the first exemplar
$s \leftarrow s'$
Repeat for each step:
  Select action $a \in A_s$ according to softmax and $Q(s, A)$ (Eqs. 2.11,5.1)
  Take action a; observe reward r and next state s'
  $\delta = r + \gamma \max_{a'} Q(s', a') - Q(s, a)$
  if $w(s, A)$ does not exist,
    $w(s, A) \leftarrow Q(s, A)$ Recruit a new exemplar
  $w(S, a) \leftarrow w(S, a) + \epsilon_w \cdot \delta \cdot \frac{sim(S,s)}{\sum sim(S,s)}$
  $\alpha_i \leftarrow \alpha_i + \epsilon_\alpha \cdot \delta \frac{\partial}{\partial \alpha_i} Q(s, a)$
  $s \leftarrow s'$

changes occur along dimensions, attention will shift to the dimensions that are most diagnostic of the correct action.

5.2 Simulation Studies with Q-ALCOVE

The added benefit of learnable selective attention in Q-ALCOVE was assessed via a GridWorld task (e.g. Sutton & Barto 1998). GridWorld tasks model an environment as a set of states arranged in a D-dimensional rectangular lattice and typically involve discovering and learning to navigate to a goal state or set of goal states. In each state, the learner has $2D$ available actions corresponding to moving in either direction along each dimension.

The task I developed to evaluate Q-ALCOVE I call Directional GridWorld because the goal states are the $(D-1)$-dimensional hyperplane that bisects the total set of states along a single dimension. This creates a task where only one dimension is relevant to the task and an arbitrary number of dimensions are irrelevant. A 3-dimensional Directional GridWorld is shown in Figure 5.1. Whenever the learner reaches a goal state, a reward of $+10$ is provided. On the next step, the learner is taken to a random state maximally distant from the goal region. All actions that do not lead to a goal incur a cost of 1 (a reward of -1). The environment within a simulation is of a fixed size, and the learner cannot leave it. Actions that would result in the learner leaving the space incur a cost of 1 and keep the learner in the same state.

Once attention is learned, the task effectively becomes one-dimensional as the learner's position along any of the irrelevant dimensions can be ignored, and only the single value describing the position along the relevant dimension is necessary to determine the correct action. The primary question being investigated with Directional GridWorld is whether Q-ALCOVE learns attention as expected in such a task: to maximally generalize across irrelevant dimensions, and to tighten its generalization gradient along the relevant dimension. This should lead to an allocation of attention weights such that the value of α for the relevant dimension is high, whereas α values for the irrelevant dimensions are low.

Three different models were run on Dimensional GridWorld to compare their performance:

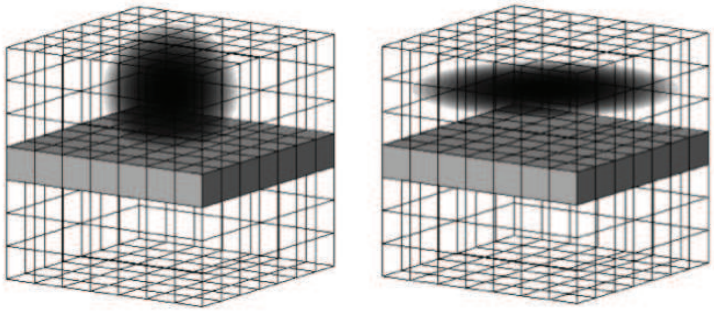

Figure 5.1: State space for 3-dimensional Directional GridWorld task. Grey states across the center are goal states. The black cloud depicts Q-ALCOVE's generalization gradient, at the start of learning (left) and after 300 time steps (right).

Q-ALCOVE itself, a version of Q-ALCOVE with the attention learning parameter $\epsilon_\alpha = 0$ so that the model's values of α would not change (this model is referred to as the fixed-generalization model), and finally a tabular Q-learning model with no generalization. Q-ALCOVE's attention learning parameter was set to $\epsilon_\alpha = .2$. All models had their other parameters equal: $\epsilon_w = .9$, and choice temperature parameter $\tau = 0.9$ (Equation 2.11). Initial values for α were all set to 0.4. This value was determined experimentally to maximize the performance of the fixed-generalization model model in three dimensions, and verified to provide good, if not maximal performance in higher dimensions. The first action in GridWorld is taken at random, and the first recruited exemplar has all its Q-values initialized to the first reward the learner receives.

Figure 5.2 shows the values for α for each of five dimensions in a single representative run of each of the three models in a 5-dimensional GridWorld. This verifies that Q-ALCOVE does effectively shift its attention to the single dimension relevant to the task. Figure 5.3 shows the average reward rate over time for each of the three models in a 5-dimensional Directional GridWorld task averaged over 100 independent runs and smoothed over a rectangular time window of 10 steps. The learning curves show that attention learning provides an advantage over the other two models. Figure 5.4 shows each model's Q-values for each action within each state in a two-dimensional slice through the three-dimensional GridWorld at time step 300. The values for the tabular Q-learning model are irregular and reflect the fact that each Q-value has been learned by direct experience. The Q-values for the fixed generalization model show that the model's predictions are more accurate, but useful values learned in one region of a row are not found uniformly across the row. Also, value estimates are inappropriately generalized across rows. In contrast, Q-ALCOVE has a highly regular and very accurate set of Q-values for the task. Their accuracy is due to the pooling of information about states from within a row in a way relevant to the task, but not making the mistake of combining information across rows. This generalization has been shaped by the learned attention parameters (Figure 5.2) changing the generalization gradient to optimize performance on the task (Figure 5.1).

To examine the performance gains afforded by learnable attention as dimensionality increases,

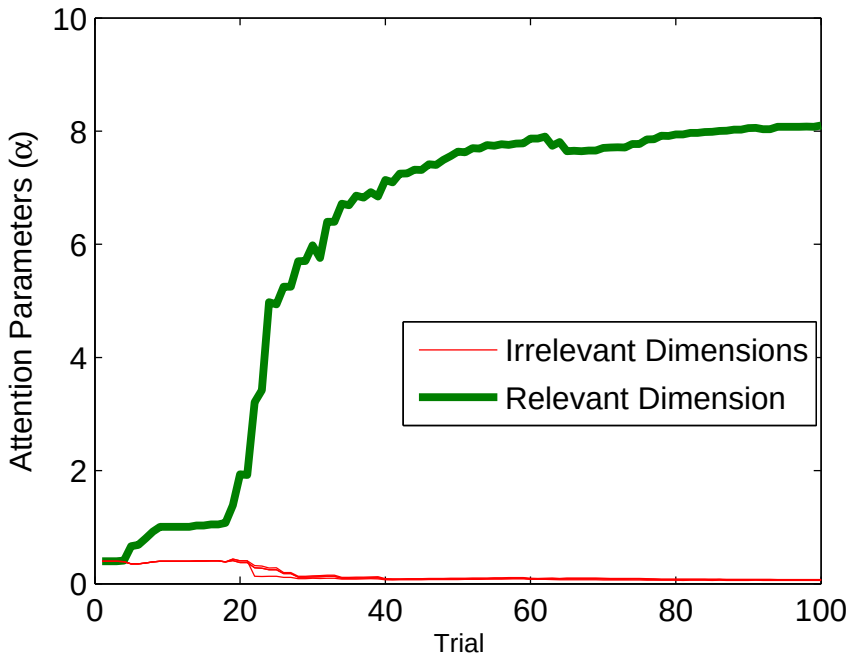

Figure 5.2: Dynamics of attention weights for one run of Q-ALCOVE in a 5-dimensional Directional GridWorld. The red trace across the bottom is actually the attention for the four irrelevant dimensions; they are disattended nearly equally over the course of learning, and so overlay each-other nearly perfectly.

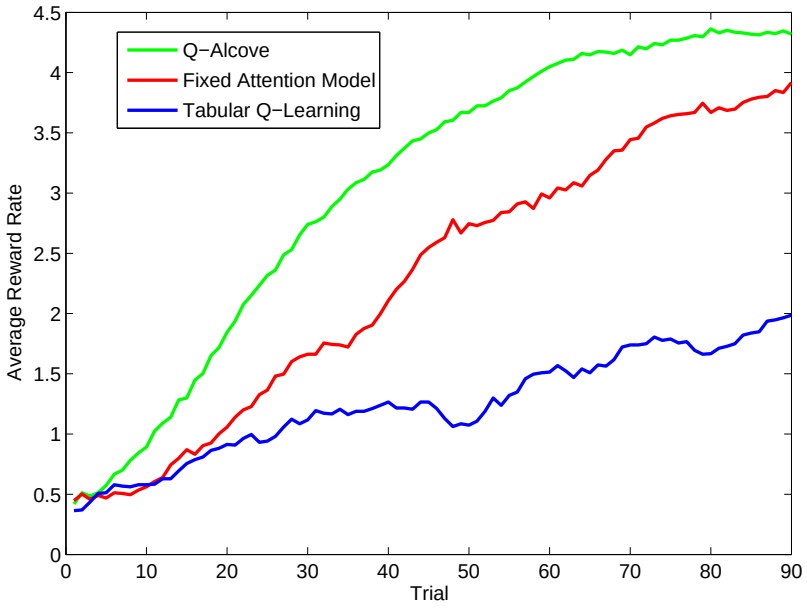

Figure 5.3: Reward rate over time in a 5-dimensional Directional GridWorld for Q-ALCOVE, the equivalent model with fixed generalization, and a tabular Q-learning model. Each learning curve is average reward for a given time step averaged over 100 trials and smoothed over a rectangular time window of 10 steps.

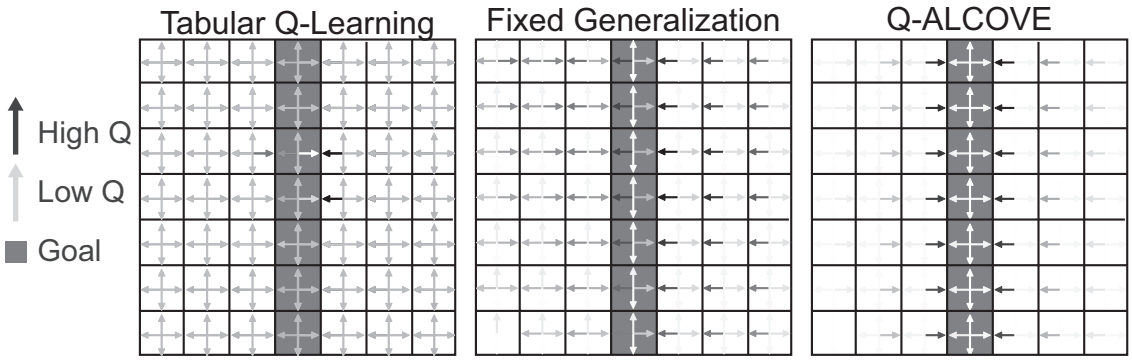

Figure 5.4: Q-values obtained from Q-ALCOVE, the equivalent model with fixed generalization, and a tabular Q-learning model after 300 steps in 3-dimensional Directional GridWorld. Shown is a 2-dimensional slice through the center of the state space. Arrows in each state correspond to the four actions within the plane. Darker arrows indicate greater Q-values.

Q-ALCOVE and the fixed-generalization model were each run in Directional GridWorlds, each three wide long each dimension, and ranging from five to fifteen dimensions. To accommodate for the large number of exemplars that would be necessary to learn in higher dimensions, Q-ALCOVE was limited to creating a maximum of 400 exemplars. Since Q-ALCOVE's learning rule updates all exemplars in proportion to their contribution to the last action, no additional accommodations were necessary for reasonable model performance. The exemplars were recruited in the normal fashion, and represented the 400 first unique states encountered. The compact nature of the 3-wide GridWorld helps ensure that worm-like paths in the state space do not unnaturally color learning. Figure 5.5 shows the learning curves for these simulations. The advantage of attention learning can be seen to increase with dimensionality.

5.3 Conflict Between Policy and Q-Values

Though Directional GridWorld was designed as a task where a model with an attentional learning mechanism would be at an advantage to a model without, it was discovered that when compared on pure reward rate, Q-ALCOVE does not always outperform its restricted counterpart with fixed generalization. The problem with Q-ALCOVE as it is presented here lies with a disagreement between optimal attention allocation to reduce TD-error and optimal attention allocation to maximize reward. Consider the second and third panels in Figure 5.4. At that point in time, the fixed generalization model's Q-values are fairly inaccurate, while Q-ALCOVE's are nearly correct. Particularly, along the right side of the shown slice of the state space, the fixed generalization model is vastly over-estimating the value of moving to the left. In this particular task that also leads to the model taking the correct action. The result is that for larger state spaces, where there are states relatively distant from the goal state, maintaining some generalization **along** the relevant dimension increases performance even when the gradient on TD-error would decrease generalization. It is only where generalization along the relevant dimension leads to an incorrect action that learning attention to minimize error on Q leads to faster learning. This is why the Attentional GridWorld was made to be three states wide along all dimensions in most of the simulations presented here. It

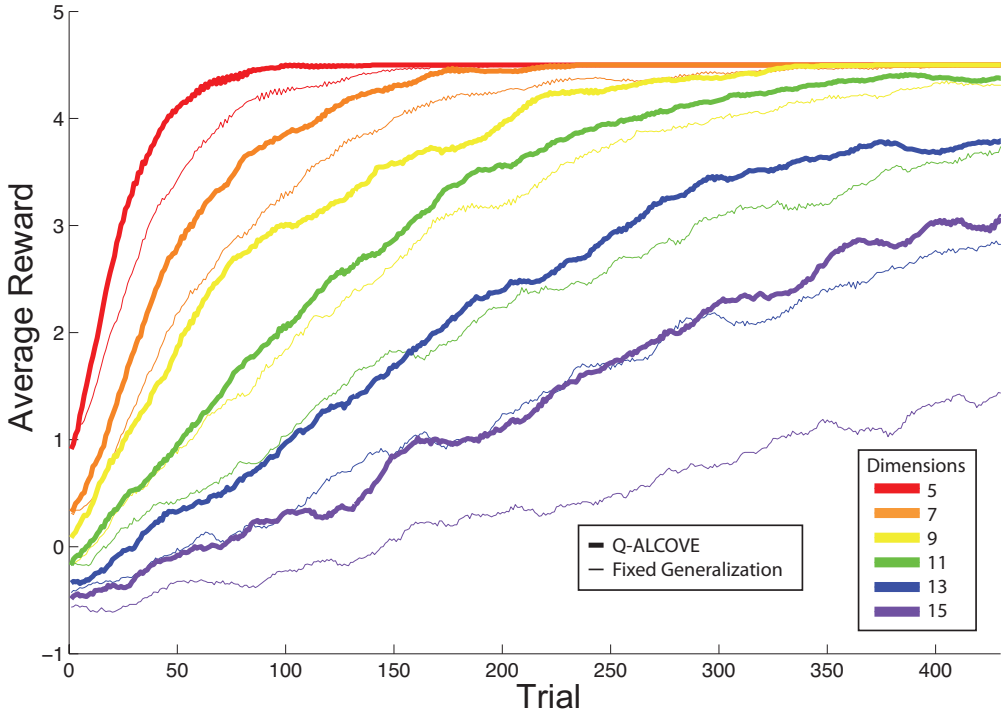

Figure 5.5: Learning curves in Directional GridWorld, 3-wide in each dimension, for Q-ALCOVE and the fixed generalization model. Each of the learning curves are the average of 70 runs and smoothed over a window of 30 steps. Note that Q-ALCOVE's advantage increases modestly with increased dimensionality.

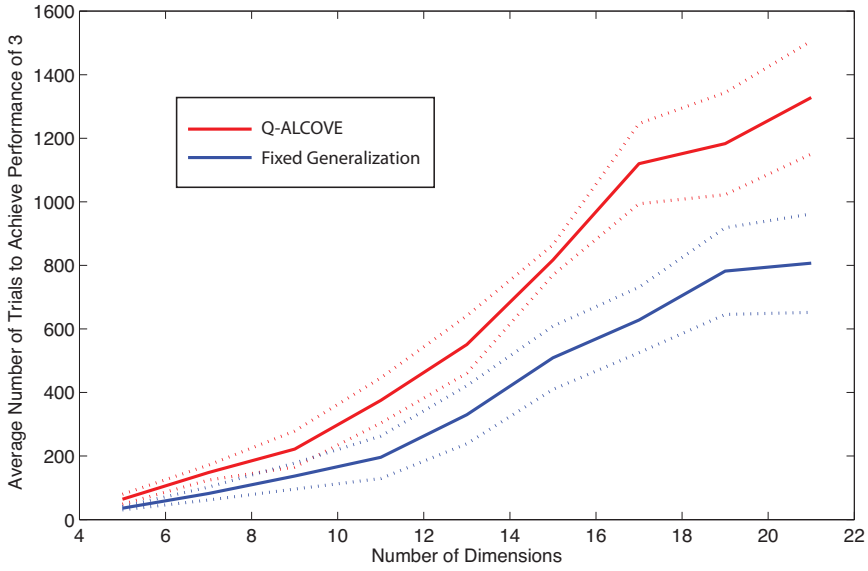

Figure 5.6: Time to reach an average reward rate of 3 in Directional GridWorld, 3-wide in each dimension, for Q-ALCOVE and the fixed generalization model. Performance was measured with a moving window of 10 trials. Each measurement is an average of the time to reach the criterion for 500 independent runs of each model.

needs to be the case that sharing generalizing along the relevant dimension leads to an **incorrect action** to have the present model outperform its more restricted form.

It is not difficult to construct other tasks where the present Q-ALCOVE will outperform a fixed generalization model. Consider a GridWorld where a single dimension is relevant to the correct action, and for every alternating layer along that dimension a different action is optimal. This leads to a situation where sharing information across layers is both not beneficial in determining Q values for the current state, but **also** leads to an incorrect action.

But the conflict between learning attention in the service of optimizing Q-values estimates versus optimal actions poses a problem for the present model. It is not sufficient to invent another category of problems to satisfy the idiosyncrasies of Q-ALCOVE. A promising direction for addressing the conflict is to move the model towards an actor-critic architecture. Actor-critic methods are differentiated from other learning architectures by a separation between their method of estimating state values (the critic) and their method of determining the correct action (the actor). But moving to an actor-critic architecture in this case is not trivial because both the actor and the critic now have separate information about the task, and so each has its own representation that may need to be addressed. From what has been observed so far in Directional GridWorld, one might guess that the critic would have an attention parameters learned via TD-error while the actor would not. But in the suggested alternating layer task, the opposite may need to be true, leading to the possibility that the actor and critic may each need to have separate pools of attention to act optimally in both kinds of tasks.

Chapter 6

Human Experiment

Q-ALCOVE's principal claim is that TD error can drive attention learning. TD error is constructed from and acts on value estimates for states (Eqs. 2.10,2.8). An investigation of these internal value estimates in the service of TD learning in humans is critical to the model's psychological validity, and to the premise that there is an interaction between RL and rich representations.

To determine whether the general approach of driving attention learning with TD error is psychologically valid, the minimum experiment should have subjects perform a two step task. The first step should present a stimulus which requires differential attentional allocations to two stimulus dimensions for optimal performance. Actions taken in the first step should provide no overt reward, and the resulting state should have no a priori value. The action taken from the second state should give the subject an overt reward, but does not necessarily need selective attention to act on it. The reward should be contingent on the second stimulus such that the optimal expected value of possible second stimuli should differ. In such a task, the overt reward after the second step is predicted to drive the learning of an internal value for the intermediate, or second stimulus. In turn, the internal value for the second stimulus is used to construct an internal covert TD-error signal, and that in turn drives attention.

In fulfillment of those requirements, I designed a task consisting of a two-step decision process in which the action on the first step probabilistically determined the stimulus on the second step. Only after the second action did the subject receive feedback about reward.

The second stage of the task was a simple decision task with two possible stimuli and two

possible actions. A different action was optimal (i.e., maximized reward) for each of these intermediate stimuli. Once this mapping was learned, one intermediate stimulus led to a higher reward than the other. RL predicts that once subjects learned the optimal actions on this second step, they would learn to assign differential values to the two intermediate stimuli. These values would in turn be used for computing a TD-error signal for actions in the first step, thereby allowing subjects to learn an action policy that maximizes the probability of obtaining the higher-valued intermediate stimulus.

The stimulus for the first choice varied on two continuous dimensions, one of which was more predictive of the outcome of the first action (i.e., the intermediate stimulus) and hence of which choice was optimal. The key question was whether learning the first action through TD error would also lead to learning of selective attention between stimulus dimensions, such that subjects would shift attention to the more relevant dimension. The stimulus set of the first step was designed so as to allow assessment of subjects' attentional allocation based on their patterns of errors, as described below.

6.1 Methods

150 undergraduate students from the University of Colorado, Boulder served as the experimental subjects in exchange for course credit.

Subjects were instructed they would pretend to be mushroom farmers. On each trial, they were presented with an image of a mushroom spore and asked to choose between two locations for growing the spore, Sun and Shade. This action determined the intermediate stimulus, a pair of blue or orange mushrooms. They were then given the option to sell the mushrooms to either a Troll or a Goblin, who paid them in gold coins. The structure of the task is outlined in Figure 6.1.

The stimulus in the first stage was a yellow spore shape, consisting of a circular center measuring 2.3 cm in diameter and radial spines arranged evenly around the center. The spines ranged from 8 mm to 260 mm in length and varied in number between 20 and 100. Spores were uniformly sampled from a circular region inscribed within this two-dimensional stimulus space.

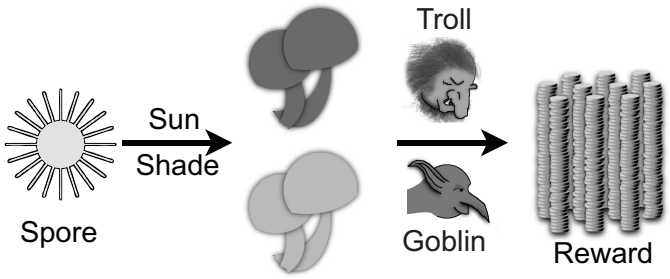

Figure 6.1: An overview of the spores task. Subjects are presented with a spore, take the first action, are presenting with a resulting mushroom, and following their second action are presented with a reward.

The spore was presented in the center of an LCD monitor over a black background. The subject selected an action by pressing either S (Sun) or C (Cave) on the keyboard. After this first response was given, the spore disappeared and a pair of cartoon mushrooms appeared in the center of the screen. The subject selected the second action by pressing T (Troll) or G (Goblin). The reward was then presented as stacks of gold coins with a numeric value underneath. The mushrooms and the chosen creature remained on the screen while the reward was displayed.

The transition after each step was animated, lasting 1200 ms between the first response and intermediate stimulus, and 970 ms between the intermediate stimulus and the reward. The reward remained on the screen for 800 ms. A blank screen separated the reward from the beginning of the next trial for 200 ms.

The reward structure for the second step was defined as shown in Table 6.1. Each mushroom color was associated with a different optimal action. Under these actions, one mushroom (henceforth referred to as the "good" mushroom) afforded a higher reward.

Table 6.1: Reward structure of the second stage of the spores task. Reward on each trial was sampled uniformly from the ranges shown.

	Creature Sold to	
Mushroom Color	Goblin	Troll
Blue	[200, 220]	[400 420]
Orange	[300, 320]	[100 120]

The transition dynamics for the first step were defined as follows. For each action, the

probability of one mushroom color versus the other was a logistic function of the dimension values of the spore, given by $p = 1/(1 + \exp(A(30L + 10N)))$, where L and N represent the length and number of the spines, scaled to range from -1 to 1, and A represents the action on the first step, coded here as ± 1. The coefficients for L and N were counterbalanced between subjects, so that L was the more relevant dimension for half the subjects and N was more relevant for the other half. The effect of this design was to create an optimal decision bound, at an angle of 18.4° to one of the two axes, such that the action that maximized the probability of obtaining the good mushroom was determined by which side of the boundary each spore lay on.

Subjects were randomly assigned to Length-relevant and Number-relevant conditions, which differed in which spore dimension was more predictive. The roles of the creatures, the colors of the mushrooms, the labels for the first action, and the direction of skew for the optimal bound were also counterbalanced between subjects. Each subject completed 240 trials (480 total decisions) in blocks of 40.

6.2 Predictions and Analysis

The theory presented the Chapter 5 predicts subjects to shift attention to the more relevant spore dimension in response to a temporal difference error based on the internal values learned for mushrooms. Under the view of attention as a transformation of perceptual space, subjects' representations of the set of spores should become stretched along the more relevant dimension and compressed along the less relevant dimension, as shown in Figure 6.2. Consider the stimuli in the highlighted areas of the figure. Under the attention-altered representation, most of their neighbors lie on the opposite side of the optimal decision bound. Therefore, similarity-based generalization will lead to higher rates of suboptimal actions for these critical stimuli, as compared to matched stimuli on the other side of the optimal bound. The same prediction arises if one assumes subjects learn prototypes for spores associated to the two actions, because each critical stimulus is more similar to the opposite prototype (taken to be the centroid of the region on that side of the optimal bound). Therefore the predictions do not depend on an assumption of exemplar-based generalization.

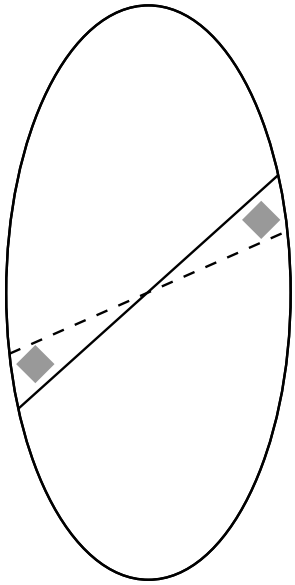

Figure 6.2: Predictions from selective attention in first step of task. Attention to the more relevant (vertical) dimension leads to stretching of the stimulus space. Critical stimuli (grey) near ends of optimal decision bound (solid line) are predicted to lead to errors as most of their neighbors lie on the opposite side of the optimal decision bound, thus producing a rotation away from optimality in the best fit of a linear classifier to subject's responses (dashed line).

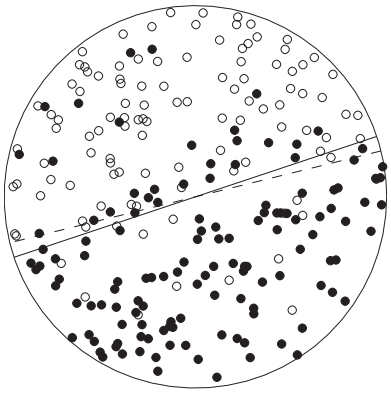

Figure 6.3: Responses on first step of spores task for a typical subject. The solid line shows the optimal bound. The dashed line shows the fit of a linear classifier.

To test this prediction, we used bivariate logistic regression to fit a linear classifier to each subject's responses. This classifier estimated a linear boundary in stimulus space that best divided the spores the subject chose to grow in the sun from those grown in the shade. To illustrate this analysis, Figure 6.3 shows the response distribution of a typical subject in the learning group (defined below). Open and closed circles represent stimuli for which the subject selected each of the two actions, the solid line represents the optimal bound, and the dashed line represents the output of the linear classifier. The prediction from selective attention, based on the analysis of expected errors described above, is that the boundary separating each subject's decisions will be rotated relative to the optimal boundary, as shown by the dashed line in Figure 6.2.

In the absence of selective attention, the representation of the stimulus space would remain circular, and therefore by symmetry there should be no systematic bias in the subject's estimated decision bound. Therefore, testing for the predicted bias is a diagnostic way to determine whether the proposed attention-learning mechanism is operating.

On average, subjects made the correct action on the second step of the task on 89.4% of trials. Figure 6.4 shows the distribution, across subjects, of the proportion of good mushrooms obtained following the first step. The histogram shows a clear bimodality, wherein many subjects performed at chance for the first step, but a significant number were able to learn effective actions.

As explained below, we only predict selective attention for subjects who learn the first stage

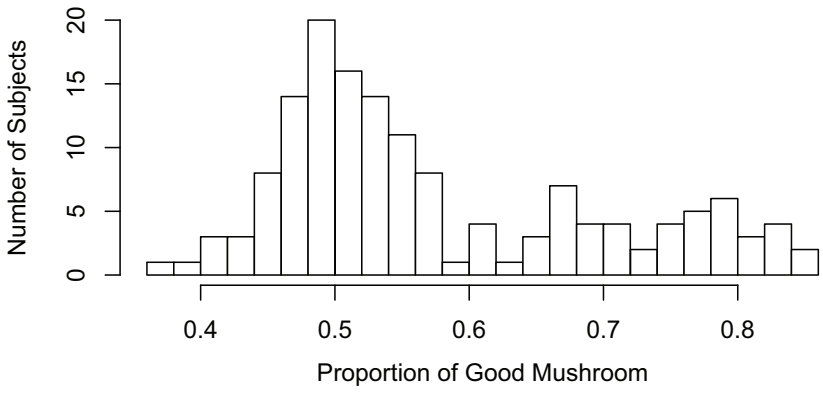

Figure 6.4: Distribution of performance on first step of spores task.

of the task. Therefore we analyzed the responses of subjects who performed above 70% on the first stage. This cutoff was based on a visual inspection of Figure 6.4 to safely exclude subjects who were performing at chance. A total of 30 subjects performed at or above 70% on the first step of the task, 11 in the Length-relevant condition and 19 in the Number-relevant condition.

A linear classifier was fit to the first-step responses of each subject in the learning group. Figure 6.5 shows the orientations of the resulting decision bounds, indicated by dots on the circumference of the stimulus region. The mean orientation for each group is shown as a dashed line, and the optimal bound as a solid line. The Number-relevant condition is shown in black and the Length-relevant condition in grey. The mean orientation of the decision bound for subjects in the Length-relevant condition was 7.96° from the Number axis. This value was significantly different from the optimal bound ($18.4°; t_{10} = -2.99, p = .014$) as well as from zero ($t_{10} = 2.29, p = .045$). The mean orientation for the Length-relevant condition was 7.33° from the Number axis. This too was significantly different from the optimal bound ($t_{18} = -3.25, p = .004$) and from zero ($t_{18} = 2.14, p = .046$).

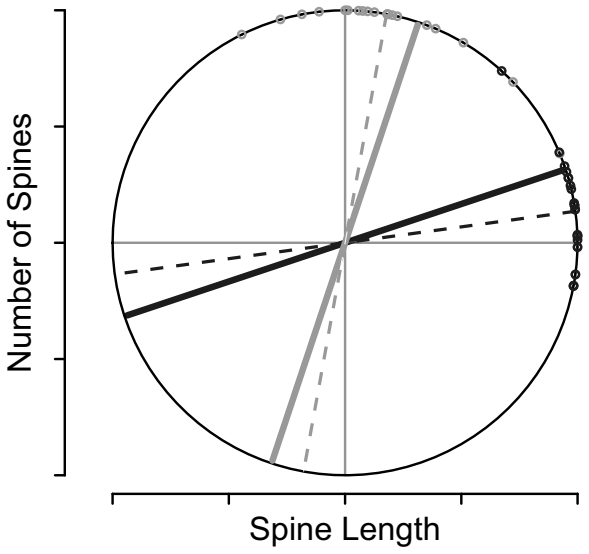

Figure 6.5: Orientations of empirical decision bounds for subjects in learning group. Small circles represent individual subjects' decision bounds; dashed lines are the groups' mean decision bounds; heavy solid lines represent the optimal bounds for each condition; black = Number-relevant; grey = Length-relevant.

6.3 Results

The results of the decision-bound analysis confirm that subjects made more errors on the critical stimuli. This prediction follows directly from the assumption of selective attention to the more relevant dimension. Because actions on the first step only led to colored mushrooms and not nominal reward, the results support the proposal that attention learning can be driven by internal value estimates and error signals.

Selective attention is only predicted for higher-performing subjects for three reasons. First, the theory of attention learning via TD-error only predicts attention to be learned once some amount of learning has taken place in associating stimuli to appropriate actions. Attention learning essentially works as a bootstrapping method operating by altering generalization and thus requires some amount of reliable knowledge to begin with in order for adaptation of generalization to have a useful effect. Second, because the theory predicts a bidirectional relationship between attention and value learning, those subjects who exhibit more selective attention should perform better on the task. Therefore, performance acts as a cue to indicate which subjects are more likely to exhibit a measurable effect. The third reason is purely methodological, in that the linear classifier requires a systematic set of responses in order to estimate a meaningful decision bound.

An alternative to the proposal of learned attention is that subjects simply disregarded one dimension of the stimulus entirely. This more strategic explanation is still consistent with the general theory of representation learning driven by RL, but the mechanism would be incompatible with continuous adjustment of attention weights. Regardless, the data rule out this explanation. The fact that the mean bound orientations were reliably different from zero (i.e., the less relevant axis) implies that subjects were sensitive to the less relevant dimension (they were just less sensitive to it than to the primary dimension). Another possibility is that some subjects disregarded one dimension and others disregarded the other, with most subjects in each condition disregarding the less relevant dimension. However, this explanation predicts a bimodal distribution of bound orientations at the subject level, which is clearly not present.

6.4 Fitting Q-ALCOVE to Subject Data

Q-ALCOVE and a nested model without attention learning (Section 5.2) were compared in their ability to fit the subject data. The maximum likelihood fit of Q-ALCOVE, and an attention-restricted model were found. The free parameters for Q-ALCOVE were an initial value for isotropic attention, Luce-choice temperature parameter (Eq. 2.11), learning rate for w (ϵ_w in Eq. 5.2), and an attention learning rate (ϵ_α in Eq. 5.3) which was fixed at zero for the inattention model.

A likelihood ratio test showed that Q-ALCOVE fit the subject data significantly better than the nested model without attention learning ($G^2 = 493.567$; $p \approx 0$). The maximum likelihood parameters for Q-ALCOVE exhibited the same bias as seen in the subject data. The restricted model did not exhibit the same bias.

Chapter 7

Conclusions

Reinforcement Learning is a powerful approach to learning. Learning driven by temporal difference error could account for learning of all sorts of information well beyond stimulus-action pairings. The approach presented here, where TD error drives representation learning, and learned representations function in the service of temporally extended tasks represents a the next logical step in the progression of learning research. Just as we can generalize the Rescorla-Wagner model of error driven learning into continuous time as TD learning (Sutton & Barto, 1990), so too should we try to explain how representations can be learned in a continuous environment without a direct teacher. We should be looking for a general architecture to drive representation learning that is more powerful and realistic than categorization error. Additionally, a relationship between the neural correlates of TD error, seated in the basal ganglia, with frontal and temporal cortical areas lends anatomical support to a rich interaction between psychological RL and richer representations.

The modeling work presented in Chapter 4 shows that with the addition of simple constructs, in this case eligibility traces, Reinforcement Learning models can sometimes exhibit behavior that is otherwise unexpected of RL. The Q-ALCOVE model presented in Chapter 5 shows a unique method of combining RL with other psychological constructs, in this case selective attention, whereby the representation serves to speed learning, but also the learning signal fundamental to RL drives the changes in representation. The experiment in Chapter 6 served to demonstrate the psychological validity of this interaction by showing that selective attention was learned in the absence of direct feedback (Section 6.3), where the selective attention improved performance on the task.

References

Alexander, G., DeLong, M., & Strick, P. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annual review of neuroscience, 9(1), 357–381.

Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98, 409–429.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. Annual Reviews in Psychology, 56, 149-178.

Bagnell, J. A., & Schneider, J. G. (2001). Autonomous helicopter control using reinforcement learning policy search methods. Proceedings of the 2001 IEEE International Conference on Robotics Automation, 1615-1620.

Barto, A. (1994). Adaptive critics and the basal ganglia. Models of information processing in the basal ganglia, 215.

Barto, A., Sutton, R., & Watkins, C. (1989). Learning and sequential decision making. Learning and computational neuroscience.

Bellman, R. (1957). Dynamic programming. Princeton, NJ: Princeton University Press.

Busemeyer, J., McDaniel, M., & Byun, E. (1997). The abstraction of intervening concepts from experience with multiple input–multiple output causal environments. Cognitive Psychology, 32(1), 1–48.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. Nature Neuroscience, 3(November), 1218–1223.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. Current Opinion in Neurobiology, 10(6), 732–739.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. Journal of Experimental Psychology: General, 127, 107-140.

Frank, M., Loughry, B., & O'REILLY, R. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. Cognitive, Affective, & Behavioral Neuroscience, 1(2), 137.

Fu, W., & Anderson, J. R. (2008). Dual learning processes in interactive skill acquisition. Journal of Experimental Psychology: Applied, 14, 179-191.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2), 155–170.

Gluck, M., & Bower, G. (1988, Jan). From conditioning to category learning: An adaptive network model. Journal of Experimental Psychology: General, 117(3), 227-247.

Holroyd, C., & Coles, M. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychological Review, 109(4), 679–708.

Jaakkola, T., Singh, S., & Jordan, M. (1995). Reinforcement learning algorithm for partially

observable markov decision problems. Advances in neural information processing systems, 345–352.

James, M., & Singh, S. (2009). SarsaLandmark: an algorithm for learning in POMDPs with landmarks. Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, 585–591.

Jones, M., Maddox, W. T., & Love, B. C. (2005). Stimulus generalization in category learning. Proceedings of the 27^{th} Annual Conference of the Cognitive Science Society, 1066-1071.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. Psychological Review, 99, 22-44.

Loch, J. (1999). The effect of eligibility traces on finding optimal memoryless policies in partially observable markov decision processes. Advances in neural information processing systems, 1010–1016.

Loch, J., & Singh, S. (1998). Using eligibility traces to find the best memoryless policy in partially observable markov decision processes. Proceedings of the Fifteenth International Conference on Machine Learning, 323–331.

Love, B. C., & Jones, M. (2006). The emergence of multiple learning systems. Proceedings of the 28^{th} Annual Conference of the Cognitive Science Society, 507-512.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. Psychological Review, 111, 309–332.

M, J., & RL, G. (2006). The structure of integral dimensions. Proceedings of the 28th Annual Conference of the Cognitive Science Society.

Medin, D., Goldstone, R., & Gentner, D. (1993, Jan). Respects for similarity (Vol. 100) (No. 2).

Medin, D., & Schaffer, M. (1978). Context theory of classification learning. Psychological Review, 85(3), 207–238.

Montague, P. R., & Dayan, P. (1998). Neurobiological modeling: squeezing top down to meet bottom up. A Companion to Cognitive Science.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. Journal of Neuroscience, 16(5), 1936-1947.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. Psychological Review, 92, 289-316.

Ng, A., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., et al. (2006). Autonomous inverted helicopter flight via reinforcement learning. In M. H. Ang & O. Khatib (Eds.), Experimental robotics IX (Vol. 21, pp. 363–372). Berlin Heidelberg: Springer.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. Journal of Experimental Psychology: Learning, 13(1), 87-108.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. Psychological Review, 101, 53–79.

O'Reilly, R., & Frank, M. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Computation, 18(2), 283–328.

Pendrith, M., & McGarity, M. (1998). An analysis of direct reinforcement learning in non-markovian domains. Proceedings of the Fifteenth International Conference on Machine Learning, 421–429.

Peng, J., & Williams, R. (1996). Incremental multi-step q-learning. Machine Learning, 22(1), 283–290.

Perez-Urihe, A., & Sanchez, E. (1999). A comparison of reinforcement learning with eligibility traces

and integrated learning, planning and reacting. Computational intelligence for modelling, control & automation: neural networks & advanced control strategies, 154.

Phillips, J., & Noelle, D. (2004, Jan). Reinforcement learning of dimensional attention for categorization. Proceedings of the Twenty-Sixth Annual Meeting of the ... . Available from http://www.cogsci.northwestern.edu/cogsci2004/papers/paper552.pdf

Poggio, T., & Girosi, F. (1998). A sparse representation for function approximation. Neural Computation, 10(6), 1445–1454.

Randløv, J., & Alstrøm, P. (1998). Learning to drive a bicycle using reinforcement learning and shaping. Proceedings of the Fifteenth International Conference on Machine Learning, 463–471.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning, II, 64-99.

Rummery, G., & Niranjan, M. (1994). On-line Q-learning using connectionist systems. Technical Report, Cambridge University Engineering Department.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80(1), 1.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. Journal of Neuroscience, 13(3), 900.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275, 1593-1599.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. Behavioral and Brain Sciences, 21, 1-54.

Sejnowski, T., Dayan, P., & Montague, P. R. (1995). Predictive hebbian learning. Proceedings of the eighth annual conference on Computational learning theory.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317–1323.

Singh, S., Jaakkola, T., & Jordan, M. (1994). Learning without state-estimation in partially observable markovian decision processes. Proceedings of the eleventh international conference on machine learning, 284–292.

Singh, S., Jaakkola, T., Littman, M., & Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. Machine Learning, 38(3), 287–308.

Smith, J., & Minda, J. (1998). Prototypes in the mist: The early epochs of category learning. Journal of Experimental Psychology: Learning, 24(6), 1411–1436.

Suematsu, N., & Hayashi, A. (1999). A reinforcement learning algorithm in partially observable environments using short-term memory. Advances in neural information processing systems, 1059–1065.

Sutherland, N., & Mackintosh, N. (1971). Mechanisms of Animal Discrimination Learning. NY: Academic Press.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. Learning and computational neuroscience: Foundations of adaptive networks, 497–537.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. The MIT Press.

Tanaka, S., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Different cortico-basal ganglia loops specialize in reward prediction on different time scales. Advances in neural information processing systems, 16.

Tesauro, G. (1995). Temporal difference learning and td-gammon. Communications of the ACM, 38(3), 58-68.

Todd, M., Niv, Y., & Cohen, J. (2009). Learning to use working memory in partially observable environments through dopaminergic reinforcement. Neural information processing systems, 1689–1696.

Watkins, C. J. C. H. (1989). Learning from delayed rewards. Unpublished doctoral dissertation, Cambridge University, Cambridge, England.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8, 279–292.